Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 162 (2019) 392–399

Procedia Computer Science

www.elsevier.com/locate/procedia

7th International Conference on Information Technology and Quantitative Management (ITQM 2019)

A cloud-based tool for sentiment analysis in reviews about restaurants on TripAdvisor

M.M. Agüero-Torales[a,1], M.J. Cobo[b], E. Herrera-Viedma[a], A.G. López-Herrera[a,1]

[a]*Dept. of Computer Science and Artificial Intelligence, University of Granada, Calle Daniel Saucedo Aranda, s/n, 18071, Granada, Spain*
[b]*Dept. Computer Science and Engineering, University of Cádiz, Avenida Ramón Puyol, 11202, Algeciras, Cádiz, Spain.*

Abstract

The tourism industry has been promoting its products and services based on the reviews that people often write on travel websites like TripAdvisor.com, Booking.com and other platforms like these. These reviews have a profound effect on the decision making process when evaluating which places to visit, such as which restaurants to book, etc.

In this contribution is presented a cloud based software tool for the massive analysis of this social media data (TripAdvisor.com). The main characteristics of the tool developed are: i) the ability to aggregate data obtained from social media; ii) the possibility of carrying out combined analyses of both people and comments; iii) the ability to detect the sense (positive, negative or neutral) in which the comments rotate, quantifying the degree to which they are positive or negative, as well as predicting behaviour patterns from this information; and iv) the ease of doing everything in the same application (data downloading, pre-processing, analysis and visualisation).

As a test and validation case, more than 33.500 revisions written in English on restaurants in the Province of Granada (Spain) were analysed.

© 2020 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)
Peer-review under responsibility of the scientific committee of the 7th International Conference on Information Technology and Quantitative Management (ITQM 2019)

Keywords: opinion mining, sentiment analysis, TripAdvisor, software tool, cloud

1. Introduction

With the advent of new ICT (Information and Communication Technologies), supported by Web 2.0, millions of people create billions of connections through the media. Each click and each keystroke creates relationships that together form a vast social network.

Users of social communication tools (email, blogs, microblogs, wikis, etc.) send fervent personal or public messages, vigorously publish opinions about a product, a person or an event, or contribute altruistically and disinterestedly to the community of knowledge to make collaborations, promote cultural heritage, or advance the cultural in the development of some product or idea. Passionate about social networks create and share (texts, images, videos, links, etc.) and value or recommend products, people and services providing help to others (whether they are neighbors or live in each other's home). extreme of the world), and expressing their

[1]Corresponding author. Tel.: +34-958-248557; fax: +34-958-243317.
E-mail address: lopez-herrera@decsai.ugr.es.

1877-0509 © 2020 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)
Peer-review under responsibility of the scientific committee of the 7th International Conference on Information Technology and Quantitative Management (ITQM 2019)
10.1016/j.procs.2019.12.002

creativity (for example, photos on Flickr.com or Instagram.com; videos on YouTube.com or Vimeo.com; etc.); thus contributing to Intelligence Web Collective. The result of all this is vast and tremendously complex networks of connections that relate people to each other, and to documents, locations, concepts, and all class of objects (mostly digital).

New opinion mining tools are now more than ever needed to collect, analyse, visualise, and generate in-depth knowledge (in the form of insights) from connection sets made up of millions of messages, links, entries, edits, photo and video updates, reviews, and product recommendations. These tools could help organisations in several ways: to know what is said (whether good or not so good) about products, services, departments of the organisation, or even the entity itself, in what sense the opinions of customers or potential consumers. To know how organisations could improve their image, products and services.

Sentiment analysis is the task of identifying and classifying the sentiments and opinions expressed in a text to understand the attitude towards a product, theme, service, etc. in particular [1]. Thus, the objective of this contribution is the presentation of a tool for opinion mining and sentiment analysis. The tool is cloud-based and focuses on the gastronomy sector and will feed from opinions published on the platform TripAdvisor.com. In order to test and validate the tool, an analysis will be made of the gastronomic context of the Province of Granada.

The rest of the contribution is structured as follows. Section Methodology summarises the main steps carried out by the tool. The proposal section displays the cloud-based technology used by the tool. Section Validation shows some of the results that the tool is able to produce. Finally, some conclusions are drawn.

2. Methodology

The methodology was divided in four steps or stages. The first was the data collection stage, next to the text preparation stage, the sentiment analysis stage with a simple rule-based model, and finally, the representation stage (evaluation and visualisation of the results). See Figure 1 (left).

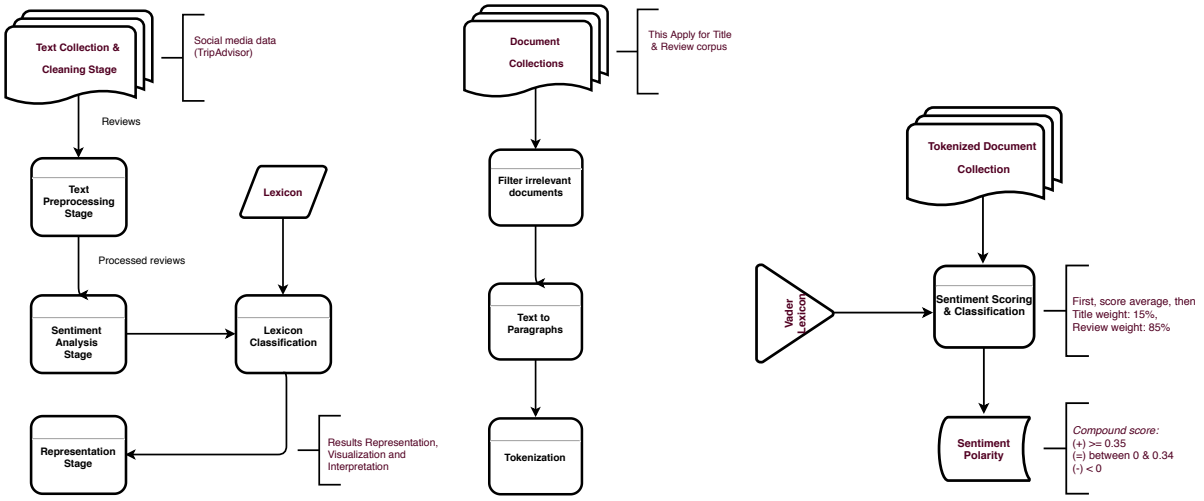

Fig. 1. Flow of the methodology. The full methodology stages (left). The text preparation stage (center). The sentiment analysis stage (right).

2.1. Data collection

We scrape TripAdvisor web-pages of all Restaurants in the Granada Province of Spain: included Bars & Pubs, Speciality Food Market, Quick Bites, Dessert, Coffee & Tea and Bakeries. According with Liu proposal [2], an opinion is a quadruple, composed by: sentiment orientation (s), sentiment target (g), opinion holder (h) and time (t). We contemplate taking some restaurants fields based on this proposal and saving them as a collection of documents (considering the following model that maps restaurants and multiple opinion relationships) [3]:

- name

- URL
- rating
- reviews collections
  - title
  - review
  - user-name
  - date
  - rating
- address
  - street
  - postal code
  - locality

As in a typical data collection stage [4], the HTML text is parsed, then scraped on TripAdvisor.com: the restaurant search URL of the Province of Granada is received, processed and navigated, retrieving all restaurant reviews (and some important data of the establishment) and, finally, reviews are recorded in a database. All reviews collected were written in English and we consider user ratings with one and two bubbles, as negative, three, as neutral, and four and five, as positive. We collected 33,594 user comments written until September 2017.

2.2. Text preparation

For text preparation stage the restaurants without reviews was discarded. The review title and the review was tokenize in paragraphs and sentences, and the text was conserved as are it, stop-words was removed but not special characters as emoticons or any others. Figure 1 (center) illustrates the text preparation stage.

In most cases on the TripAdvisor reviews, it is rare to see links next to the text, maybe emoticons, but for the sentiment analysis we use a emoticons corpus to treat it. Basically, a text analysis is performed without a very careful data cleansing strategy, based on valence[1], where the intensity of the feeling is taken into account. In turn, this type of sentiment analysis is based on word lexicons (and emoticons). By using this approach, each word in the lexicon is classified as to whether it is positive or negative, almost always how positive or negative it is [6].

2.3. Sentiment analysis

The different levels of sentiment analysis are the document level, to classify whether a whole opinion document expresses a positive or negative sentiment [7] [8], the sentence level, to determine whether each sentence expresses a positive, negative, or neutral opinion. Finally, the aspect level, to discover the target of opinion: what people like and dislike exactly [2].

In Figure 1 (right) you can see the stage of sentiment analysis, its check if any of the words in the sentences are present in the lexicon, the input text produces an output of four feeling metrics from these word ratings, which contains the proportion of negative, positive and neutral words of the given text, and a compound value, with values between -1 and 1. The compound value determines the polarity of the text, is a sum of all the qualifications of the lexicon applied [6].

2.4. Representation and web interface

The representation stage exposes what is interesting about the data, which covers the results representation and reporting, their visualisation and interpretation to understand them [1] [9].

This stage implies a responsive design for the web interface with a menu of options, also a list of restaurants. Figure 2 (top) illustrates the home screen with statistics.

In Figure 2 (bottom-left), the restaurant details screen, the user can see the polarity of sentiments of all the reviews and one-by-one, such as an alternative weighting factor to TripAdvisor bubbles, also a link to TripAdvisor.com to corroborate the sources, the same capabilities have the user details screen (see Figure 2 (bottom-right)),

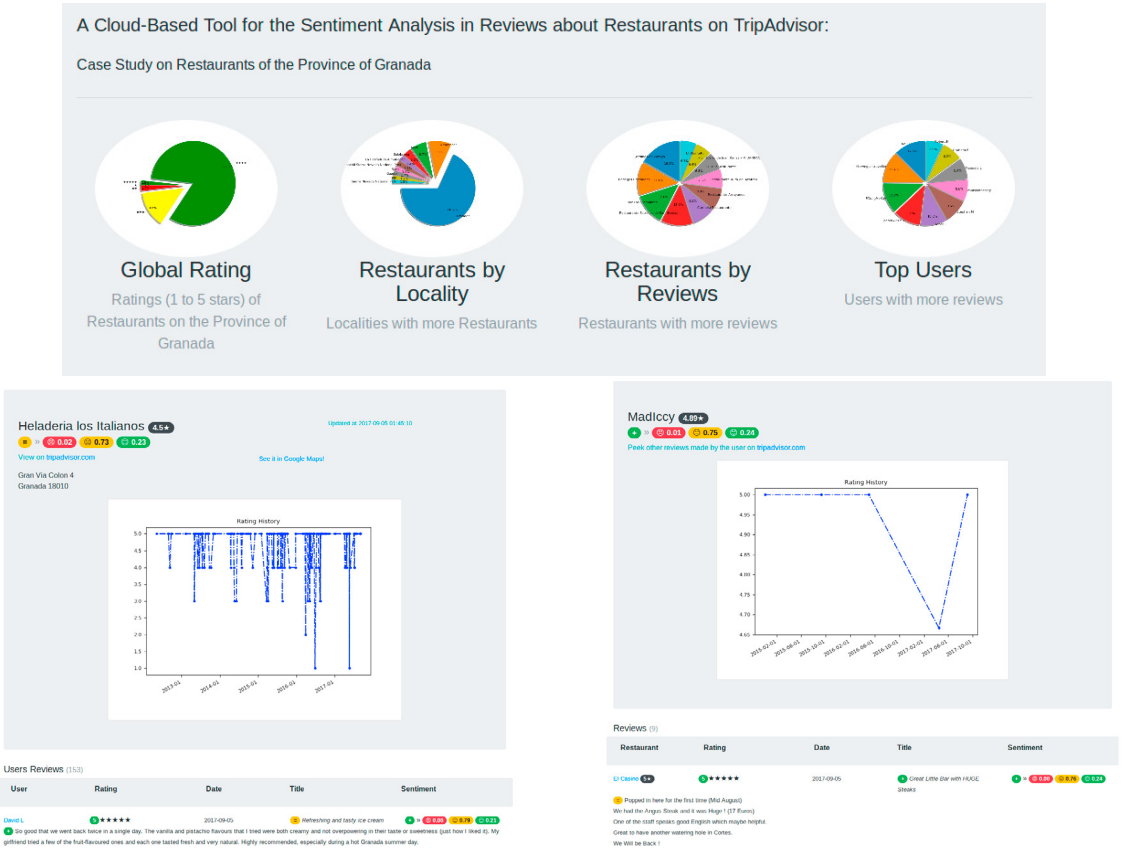

Fig. 2. Kind of analysis possibilities: general, restaurants, users.

which shows the overall polarity of the user, but restaurant screen has a link to Google Maps to see more extra comments and information.

Other important features are the on-demand TripAdvisor scraper based on an HTTP server that can control scraping through HTTP requests, the restaurants simple search engine, the user manager for authentication and registration, a REST API from the restaurant collection to load more data or download existent ones, and the calculation of the sentiment polarity on request for a particular analysis.

2.5. *Software tool*

Modern architectures have as a common general objective to seek consistency in the speed of response to the user, the use of own resources combined with third-party resources, and are based on agile development methodologies and continuous integration and continuous deployment; unlike a monolithic one, which is not suitable for modern applications because it has a single client running the user interface and a single server (replicated or not) running all the components of that application, neither because of its scalability characteristics, nor for the distribution of tasks and data between different parts of a development team [10].

To create the software tool for mass analysis of social media data, container technology was used. Containers are virtualized at the operating system level, while hypervisor-based solutions are virtualized at the hardware level, therefore, for a container, resource utilization is much more efficient, isolated and cheaper (even both technologies can complement each other for this reason) [11].

[1] Emotional valence is a psychology term refers to the intrinsic attractiveness (good-ness, positive valence) or averseness (bad-ness, negative valence) of an event, object, or situation, especially used when discussing emotions [5].

In addition of the container technology, a virtualized web environment were be created accessible from any device from a browser, as it does computing in the cloud for modern applications (see Figure 3). Then, instead of having a single service, by applying a micro-service view, each container becomes a service by itself, and the services communicate with each other through calls, gaining remarkable scalability.

Fig. 3. Web architecture and deployment diagram [12].

3. Proposal

The software tool were developed with a Python stack [1], support by Scrapy, a web crawling framework for web scraping [13], NLTK (Natural Language ToolKit), a toolkit for Natural Language Processing (NLP) provides interfaces to over 50 corpora and lexical resources , along with a suite of text processing libraries (e.g., classification, tokenization) [14], Matplotlib, a 2D plotting library for data visualisation [15], and Django[2], a potent web framework with many plugins that encourages rapid development.

The web environment relies in the virtualisation technology, focus on automation with Vagrant, a tool for building and managing virtual machine environments in a easy-to-use and single workflow [16], VirtualBox[3], a cross-platform virtualization application, a robust Docker stack, for harnessing the benefits of containerization for a focused purpose (i.e., the lightweight packaging and deployment of applications) [11], and for the data persistence, MongoDB, a general purpose, document-based, distributed database [3], where each restaurant is a document. See the Figure 4.

The sentiment analysis stage relied in VADER (Valence Aware Dictionary for sEntiment Reasoning), a lexicon and rule-based sentiment analysis tool with an overall precision of 99% for tweets sentiment classification, this tool is embedded in NLTK as a module, which proportionate other tools for text preparation and cleaning, all-in-one, but VADER no need many text preparation, it incorporate word-order sensitive relationships between terms for punctuation (e.g.,!), capitalization (e.g., ALL-CAPS), degree modifiers, support for the contrastive conjunction "but", tri-grams preceding a sentiment-laden lexical feature, or emoticons, etc. [6]. In this pipeline stage, the title and the text of the reviews is prepare and send to VADER as a sentence, the title has a weight of 15% and the review has a 85% for determine the overall polarity of the review. In a paragraph, each sentence is calculated with VADER and are calculated one-to-one for take the average of each one. This average of the compound scores (between -1 and 1) determine the final polarity, that is positive, if average equals or is major to 0.35, neutral, if average is between 0 and 0.34, and negative, if minor to 0.

4. Validation

In order to test the tool, some analyses were carried out. For example, ten best restaurants in the Province of Granada were brought together according to the ratings (date: September 13th, 2017) of TripAdvisor users and the number of revisions (see Table 1). According to this ranking, in that date the best restaurant was "El Mercader"

[2]https://www.djangoproject.com/foundation/
[3]https://www.virtualbox.org/wiki/VirtualBox

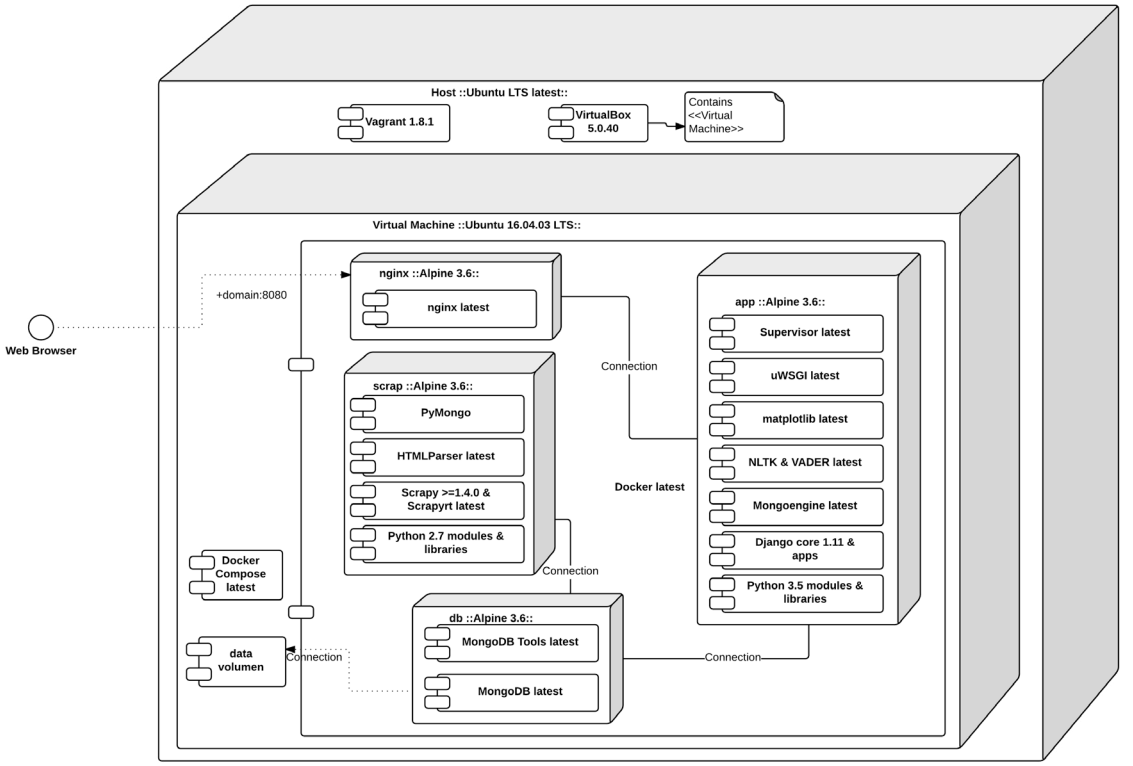

Fig. 4. General architecture deployed.

(Granada) and the tenth was "Restaurante Las Chimeneas" (Mairena, Nevada). For these restaurants the sentiment polarity was individually calculated by the software tool, and the results can be seen in Figure 5. It can be seen how the restaurant "Duran Barista" was the one that added more positivity among the reviews that were made about it. There were hardly any negative ratings. Almost the majority of the assessments were neutral.

Table 1. The ten best restaurants in the Province of Granada by TripAdvisor user rating and number of reviews. Date: September 13th, 2017.

	Restaurant	Stars	Compound	Negative	Neutral	Positive	Locality	Reviews
1	El Mercader	5	0,45	0,01	0,67	0,31	Granada	152
2	Restaurante Casa Piolas	5	0,39	0,01	0,73	0,26	Algarinejo	70
3	Duran Barista	5	0,46	0,02	0,65	0,34	Granada	40
4	Dulcimena Coffee & Go	5	0,44	0,02	0,68	0,30	Granada	32
5	D'eti Coffee And Cake	5	0,51	0,01	0,66	0,32	Granada	32
6	Vega - Foodie Bar	5	0,48	0,01	0,67	0,30	Granada	28
7	La Huella Gastrobar	5	0,48	0,01	0,64	0,33	Algarinejo	27
8	Colagallo Craft Beers & Cocktails	5	0,51	0,01	0,65	0,32	Granada	24
9	Eco De.leite	5	0,42	0,04	0,68	0,28	Granada	14
10	Restaurante Las Chimeneas	5	0,46	0,02	0,68	0,28	Mairena	13

By way of illustration, some other results that the tool is capable of producing are shown in the Figure 6. The overall rating on restaurants in the Province of Granada is shown at the top-left. Restaurants with the highest number of reviews in the Province of Granada are shown at the top-right. The locations with more establishments in the Province of Granada can be seen at the bottom-left. And finally, the bottom-right of the Figure 6 shows the Users with the highest number of restaurant reviews.

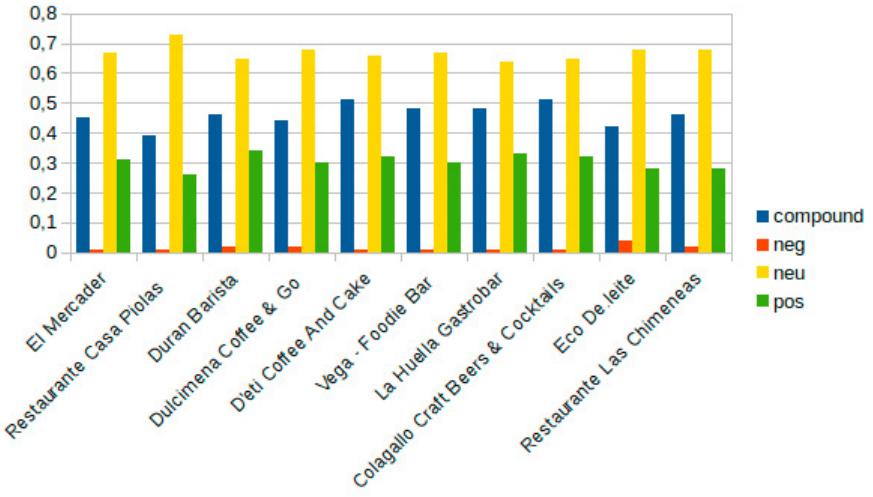

Fig. 5. Results about the ten first restaurants of Granada.

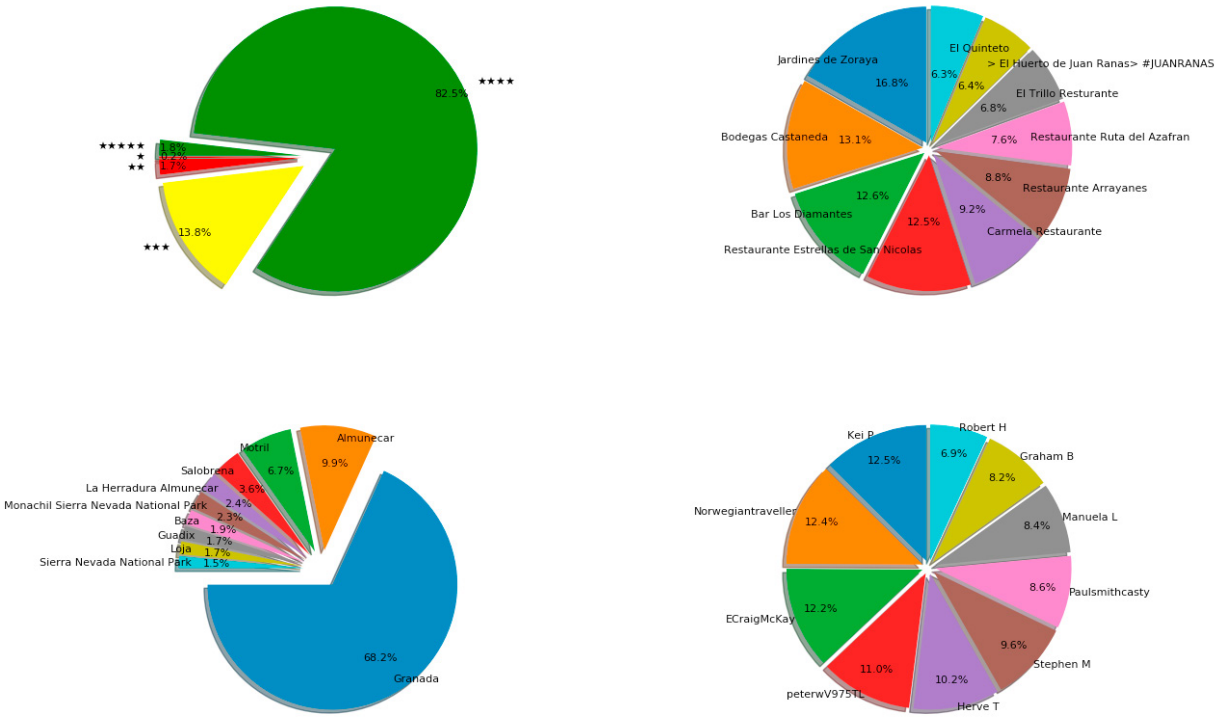

Fig. 6. Some other results. Overall rating on restaurants in the Province of Granada (top-left). Restaurants with the highest number of reviews in the Province of Granada (top-right). Locations with more establishments in the Province of Granada (bottom-left). Users with the highest number of restaurant reviews in the Province of Granada (bottom-right).

5. Conclusion

TripAdvisor is a social media par excellence to share opinions on sites related to travel, such as gastronomic establishments, that's why it has thousands of reviews a day.

With this in mind, the purpose of this contribution is to present a tool for analyzing the sentiments of gastronomic establishments. The tool has been developed using cloud-based technologies and allows all the necessary steps to carry out this type of analysis, from downloading data to displaying results, without forgetting all the intermediate stages of pre-processing, cleaning, data preparation, dimensionality reduction, etc.

The tool has been successfully validated in the gastronomic context of the Province of Granada (Spain). Some results have been shown in this contribution.

As future works we plan the improvement of some components of the tool, such as the integration with other data sources of specific purpose as Booking.com, and others of general purpose as Twitter, Instagram, YouTube, etc.

We also plan to delve deeper into online sentiment analysis methods such as MeaningCloud[4], Indico.Io[5] or Google Cloud[6], as well as explore other approaches such as aspect analysis.

References

[1] M. Bonzanini, Mastering social media mining with Python, Packt Publishing Ltd, 2016.
[2] B. Liu, Sentiment analysis: Mining opinions, sentiments, and emotions, Cambridge University Press, 2015.
[3] K. Banker, MongoDB in action, Manning Publications Co., 2011.
[4] R. Mitchell, Web Scraping with Python: Collecting More Data from the Modern Web, " O'Reilly Media, Inc.", 2018.
[5] N. H. Frijda, The Emotions, Cambridge University Press, 1986.
[6] C. J. Hutto, E. Gilbert, Vader: A parsimonious rule-based model for sentiment analysis of social media text, in: Eight International AAAI conference on weblogs and social media, 2014.
[7] B. Pang, L. Lee, S. Vaithyanathan, Thumbs up?: sentiment classification using machine learning techniques, in: Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, Association for Computational Linguistics, 2002, pp. 79–86.
[8] P. D. Turney, Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews, in: Proceedings of the 40th annual meeting on association for computational linguistics, Association for Computational Linguistics, 2002, pp. 417–424.
[9] D. A. Keim, Information visualization and visual data mining, IEEE transactions on Visualization and Computer Graphics 8 (1) (2002) 1–8.
[10] J. J. Merelo, University of Granada, Lecture notes in Cloud Computing: Arquitecturas software para la nube (September 2019). URL `https://jj.github.io/CC/`
[11] D. Merkel, Docker: lightweight linux containers for consistent development and deployment, Linux Journal 2014 (239) (2014) 2.
[12] J. M. Guirao, University of Granada, Lecture notes in Sistemas Software Basados en Web: Preliminares (September 2017). URL `https://swad.ugr.es/es`
[13] D. Myers, J. W. McGuffee, Choosing scrapy, Journal of Computing Sciences in Colleges 31 (1) (2015) 83–89.
[14] E. Loper, S. Bird, Nltk: the natural language toolkit, arXiv preprint cs/0205028.
[15] A. Devert, matplotlib Plotting Cookbook, Packt Publishing Ltd, 2014.
[16] M. Hashimoto, Vagrant: up and running: create and manage virtualized development environments, " O'Reilly Media, Inc.", 2013.

[4] https://www.meaningcloud.com/developer/sentiment-analysis
[5] https://www.indico.io/blog/docs/indico-api
[6] https://cloud.google.com/natural-language/docs/